RESEARCH ARTICLE SUMMARY

PROTEIN DESIGN

Generalized biomolecular modeling and design with RoseTTAFold All-Atom

Rohith Krishna†, Jue Wang†, Woody Ahern†, Pascal Sturmfels, Preetham Venkatesh‡, Indrek Kalvet‡, Gyu Rie Lee‡, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, David Baker*

INTRODUCTION: Proteins rarely act alone; they form complexes with other proteins in cell signaling, interact with DNA and RNA during transcription and translation, and interact with small molecules both covalently and noncovalently during metabolism and signaling. Despite substantial recent progress in protein-only structure prediction, modeling such general biomolecular assemblies remains an outstanding challenge.

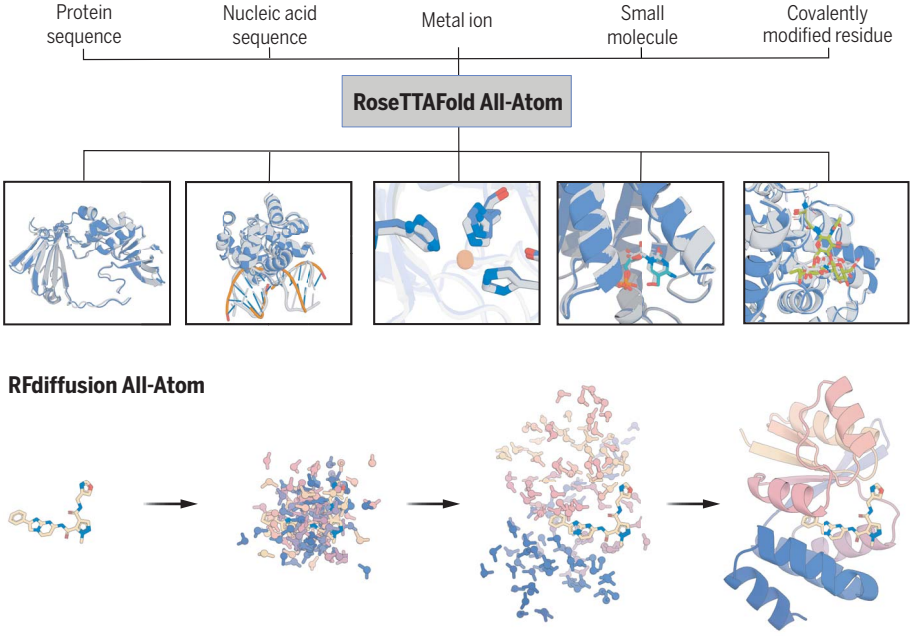
RATIONALE: We set out to develop a unified structure prediction and design approach for assemblies containing proteins, nucleic acids, small molecules, metals, and chemical modifications. We sought to combine a sequence-based description of proteins and nucleic acids with an atomic graph representation of small molecules and protein covalent modifications. We started with the RoseTTAFold2 (RF2) network, which takes as input one-dimensional (1D) sequence information, 2D pairwise distance information from homologous templates, and 3D coordinate information and iteratively improves predicted structures through many hidden layers.

RESULTS: For our biomolecular structure prediction network RoseTTAFold All-Atom (RFAA) (see the figure, top), we retained the representations of protein and nucleic acid chains from RF2 and represented arbitrary small molecules as atom-bond graphs. To the 1D track, we input the chemical element type of each nonpolymer atom; to the 2D track, the chemical bonds between atoms; and to the 3D track, information on chirality. Immediately after input, the full system was represented as a disconnected gas of amino acid residues, nucleic acid bases, and freely moving atoms, which was successively transformed through the many blocks of the network into physically plausible assembly structures. We trained RFAA on protein–small molecule, protein-metal, and covalently modified protein complexes that are found in the Protein Data Bank (PDB), filte… out common solvents and crystallization addi…

In the CAMEO blind ligand-docking challenge, RFAA outperforms baseline automated pipelines. Although not all predictions are accurate, the network outputs a confidence estimate that correlates with accuracy. The network generalizes beyond the training set: Accurate predictions are made for proteins with low sequence homology (BLAST e-value >1) and ligands with low similarity (Tanimoto similarity <0.5) to those in the training set. Prediction accuracy is higher for protein–small molecule complexes with more-favorable computed interaction energies using a molecular mechanics force field, which suggests that RFAA has learned aspects of the physical chemistry of protein–small molecule interactions. Nearly half (46%) of covalent modifications are predicted accurately [<2.5-Å root mean square deviation (RMSD)]. These additional prediction capabilities do not come at the expense of the protein structure–prediction task because RFAA has a prediction accuracy on protein monomer structures comparable to that AlphaFold2.

For small-molecule binder design, we developed RFdiffusion All-Atom (RFdiffusionAA) by fine-tuning RFAA on diffusion denoising tasks. Starting from random residue distributions, RFdiffusionAA generates folded protein structures that surround the small molecule. In contrast to previous approaches that rely on native or preexisting designed scaffolds, the binding pockets are custom generated for each ligand of interest. We generated small-molecule binding designs for the cardiac disease drug digoxigenin, the enzyme cofactor heme, and optically active bilin molecules. In each case, experimental characterization showed that a subset of the designs had the designed binding activity. The crystal structure of a heme binding design matched the RFdiffusionAA model very closely (0.86-Å Cα RMSD).

CONCLUSION: RFAA demonstrates that a single neural network can be trained to accurately model general biomolecular assemblies containing a wide diversity of nonprotein components. Although there is still room for improvement in prediction accuracy, we anticipate that RFAA should be broadly useful for modeling full biological assemblies and RFdiffusionAA for designing small molecule–binding proteins and sensors. ■



Prediction and design of biomolecular assemblies. RFAA enables the prediction of biomolecular assemblies, including proteins, nucleic acids, metals, small molecules, and covalent modifications (top). RFdiffusionAA builds de novo proteins around small molecules of interest (bottom).

The list of author affiliations is available in the full article online.
***Corresponding author. Email: dabaker@uw.edu**
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as R. Krishna *et al.*, *Science* 384, eadl2528 (2024). DOI: 10.1126/science.adl2528

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adl2528

RESEARCH ARTICLE

PROTEIN DESIGN

Generalized biomolecular modeling and design with RoseTTAFold All-Atom

Rohith Krishna[1,2]†, Jue Wang[1,2]†, Woody Ahern[1,2,3]†, Pascal Sturmfels[1,2,3], Preetham Venkatesh[1,2,4]‡, Indrek Kalvet[1,2,5]‡, Gyu Rie Lee[1,2,5]‡, Felix S. Morey-Burrows[6], Ivan Anishchenko[1,2], Ian R. Humphreys[1,2], Ryan McHugh[1,2,4], Dionne Vafeados[1,2], Xinting Li[1,2], George A. Sutherland[6], Andrew Hitchcock[6], C. Neil Hunter[6], Alex Kang[2], Evans Brackenbrough[2], Asim K. Bera[2], Minkyung Baek[7], Frank DiMaio[1,2], David Baker[1,2,5]*

Deep-learning methods have revolutionized protein structure prediction and design but are presently limited to protein-only systems. We describe RoseTTAFold All-Atom (RFAA), which combines a residue-based representation of amino acids and DNA bases with an atomic representation of all other groups to model assemblies that contain proteins, nucleic acids, small molecules, metals, and covalent modifications, given their sequences and chemical structures. By fine-tuning on denoising tasks, we developed RFdiffusion All-Atom (RFdiffusionAA), which builds protein structures around small molecules. Starting from random distributions of amino acid residues surrounding target small molecules, we designed and experimentally validated, through crystallography and binding measurements, proteins that bind the cardiac disease therapeutic digoxigenin, the enzymatic cofactor heme, and the light-harvesting molecule bilin.

The deep neural networks AlphaFold2 (AF2) (*1*) and RoseTTAFold (RF) (*2*) enable high-accuracy prediction of protein structures from amino acid sequences. However, in nature, proteins rarely act alone; they form complexes with other proteins in cell signaling, interact with DNA and RNA during transcription and translation, and interact with small molecules, both covalently and noncovalently, during metabolism and signaling. Modeling such general biomolecular assemblies composed of polypeptide chains, covalently modified amino acids, nucleic acid chains, and arbitrary small molecules remains an outstanding challenge. One approach is to model the protein chains using AF2 or RF and then successively add in the nonprotein components using classical docking methods (*3*–*9*); however, systematically evaluating and optimizing such procedures is not straightforward. RF has been extended to model both protein and nucleic acids by increasing the size of the residue alphabet to 28 (20 amino acids, four DNA bases, and four RNA bases) with RoseTTAFold nucleic acid (RFNA) (*10*), but general biomolecular system modeling is a more challenging problem, given the great diversity of possible small-molecule components. An approach capable of accurately predicting the three-dimensional (3D) structures of biomolecular assemblies starting only from knowledge of the constituent molecules (and not their 3D structures) would have a broad impact on structural biology and drug discovery and open the door to deep learning–based design of protein–small molecule assemblies.

We set out to develop a structure prediction method capable of generating 3D coordinates for all atoms of a biological unit, including proteins, nucleic acids, small molecules, metals, and chemical modifications (Fig. 1A). The first obstacle we faced in taking on the broader challenge of generalized biomolecular system modeling was how to represent the components. Existing protein structure prediction networks represent proteins as linear chains of amino acids, and this representation can be readily extended to nucleic acids. However, many of the small molecules that proteins interact with are not polymers, and it is unclear how to model them as a linear sequence. A natural way to represent the bonded structure of small molecules is as graphs whose nodes are atoms and whose edges represent bond connectivity. This graph representation is not suitable for proteins because they contain many thousands of atoms; hence, modeling whole proteins at the atomic level is computationally intractable. To overcome this limitation, we sought to combine a sequence-based description of biopolymers (proteins and nucleic acids) with an atomic graph representation of small molecules and protein covalent modifications.

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98105, USA. [3]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA. [4]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA. [5]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA. [6]School of Biosciences, University of Sheffield, Sheffield S10 2TN, UK. [7]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea.
*Corresponding author. Email: dabaker@uw.edu
†These authors contributed equally to this work.
‡These authors contributed equally to this work.

Generalizing structure prediction to all biomolecules

We modeled the network architecture after the RoseTTAFold2 (RF2) protein structure prediction network, which accepts 1D sequence information, 2D pairwise distance information from homologous templates, and 3D coordinate information and iteratively improves predicted structures through many hidden layers (*11*). We retained the representations of protein and nucleic acid chains from RF2 and represented arbitrary small molecules, covalent modifications, and unnatural amino acids as atom-bond graphs. To the 1D track, we input the chemical element type of each nonpolymer atom; to the 2D track, the chemical bonds between atoms; and to the 3D track, information on chirality [whether chiral centers are (*R*) or (*S*)]. For the 1D track, we supplemented the 20-residue and 8–nucleic acid base representation in RFNA with 46 new element-type tokens representing the most common element types found in the Protein Data Bank (PDB) (table S5). For the 2D track atom-bond embedding, we encoded pairwise information about whether bonds between pairs of atoms are single, double, triple, or aromatic. These features are linearly embedded and summed with the initial pair features at the beginning of every recycle of the network, which allows the network to learn about bond lengths, angles, and planarity. Because the 1D and 2D representations in the network are invariant to reflections, we encoded stereochemistry information in the third track by specifying the sign of angles between the atoms surrounding each chiral center (fig. S1); at each block in the 3D track, the gradient of the deviation of the actual angles from the ideal values (with respect to the current coordinates) was computed and provided as an input feature to the subsequent block (Fig. 1B).

Unlike proteins and nucleic acid sequences, molecular graphs are permutation invariant, and hence, the network should make the same prediction irrespective of small-molecule element token order. In AF2 and RF2, the sequence order of amino acids and bases is represented by a relative position encoding; for atoms, we omitted such an encoding and leveraged the permutation invariance of the network attention mechanisms. We also modified the coordinate updates: In AF2 and RF, protein residues are represented by the coordinates of the Cα and the orientation of the N-C-C rigid frame (or the P coordinate and the OP1-P-OP2 frame orientation in RFNA), and, along the 3D track, the network generates rotational updates to each frame orientation and translational updates to each coordinate. To generalize this representation in our biomolecular structure prediction network RoseTTAFold All-Atom (RFAA), heavy-atom coordinates are added to the 3D track and move independently based only on a predicted

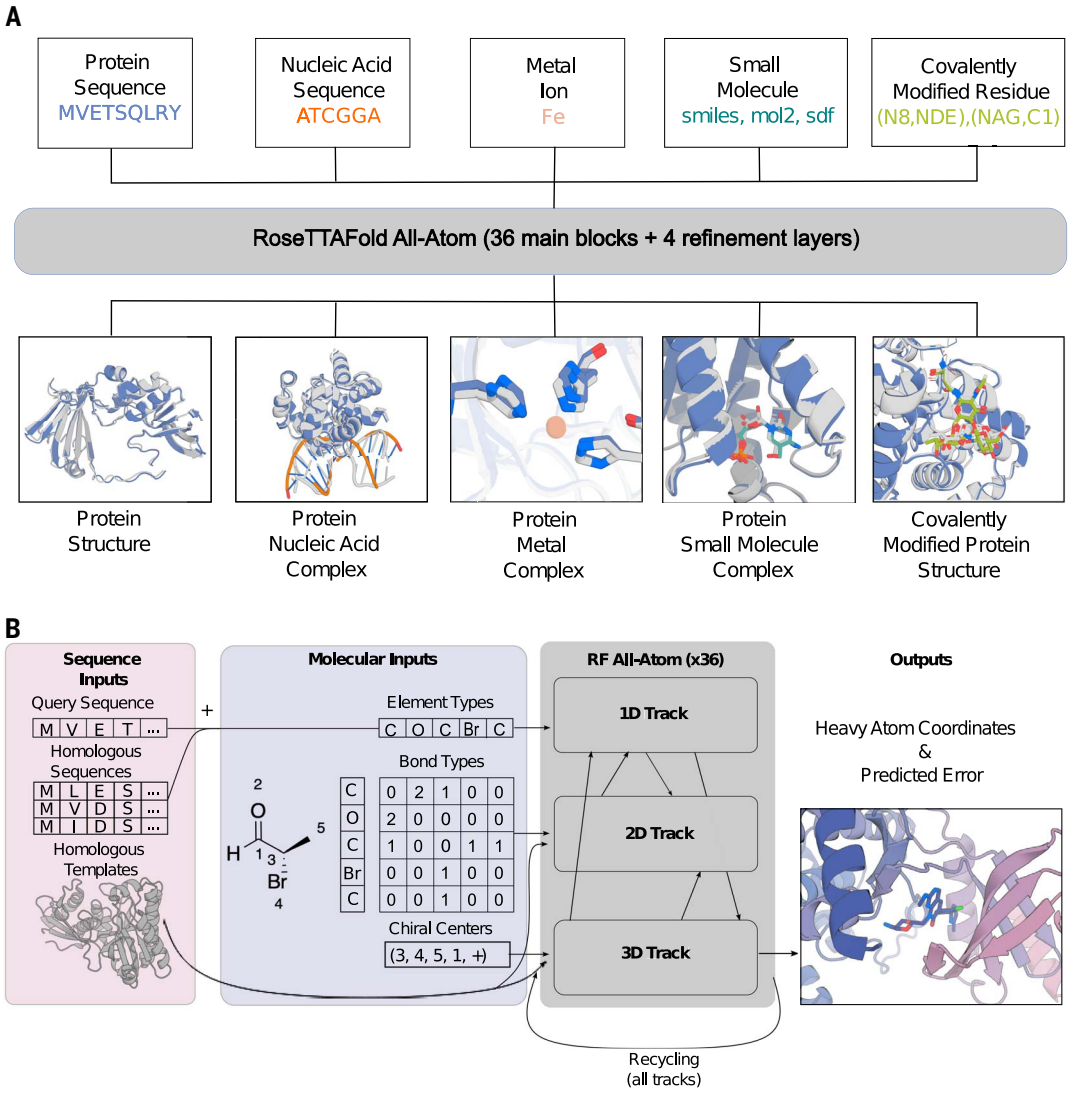


Fig. 1. General biomolecular modeling with RFAA. (**A**) RFAA takes input information about the molecular composition of the biomolecular assembly to be modeled, including protein amino acid and nucleic acid base sequences, metal ions, small molecule–bonded structure, and covalent bonds between small molecules and proteins. (**B**) Processing of molecular input information. Small-molecule information is parsed into element types (46 possible types), bond types, and chiral centers. Covalent bonds between proteins and small molecules are provided as a separate token in the bond adjacency matrix. The three-track architecture mixes 1D, 2D, and 3D information and predicts all-atom coordinates and model confidence. Single-letter abbreviations for amino acid residues in the figures are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

translational update to their position. Thus, immediately after input, the full system is represented as a disconnected gas of amino acid residues, nucleic acid bases, and freely moving atoms, which is successively transformed through the many blocks of the network into physically plausible assembly structures. For the loss function to guide parameter optimization, we developed an all-atom version of the frame-aligned point error (FAPE) loss introduced in AF2 by defining coordinate frames for each atom in an arbitrary molecule based on the identities of its bonded neighbors and, as with residue-based FAPE, successively aligning each coordinate frame and computing the coordinate error on the surrounding atoms (Fig. 2A; for greater sensitivity to small-molecule geometry, we upweighted contributions involving atoms; see supplementary methods). In addition to atomic coordinates, the network predicts atom and residue-wise confidence [predicted local distance difference test (pLDDT)] and pairwise confidence [predicted alignment error (PAE)] metrics to enable users to identify high-quality predictions. A full description of the RFAA architecture is provided in the supplementary methods.

Training RFAA

We curated a protein-biomolecule dataset from the PDB that includes protein–small molecule, protein-metal, and covalently modified protein complexes, filtering out common solvents and crystallization additives. After clustering (30% sequence identity) to avoid bias toward overrepresented structures, we obtained 121,800 protein–small molecule structures in 5662 clusters, 112,546 protein-metal complexes in 5324 clusters, and 12,689 structures with covalently modified amino acids in 1099 clusters for training. To help the network learn the general properties of small molecules rather than features specific to the molecules in the PDB, we supplemented the training set with small-molecule crystal structures from the Cambridge Structural Database (*12*). Each training example is sampled uniformly from the set of organic nonpolymeric molecules, and the network predicts the coordinates for the asymmetric

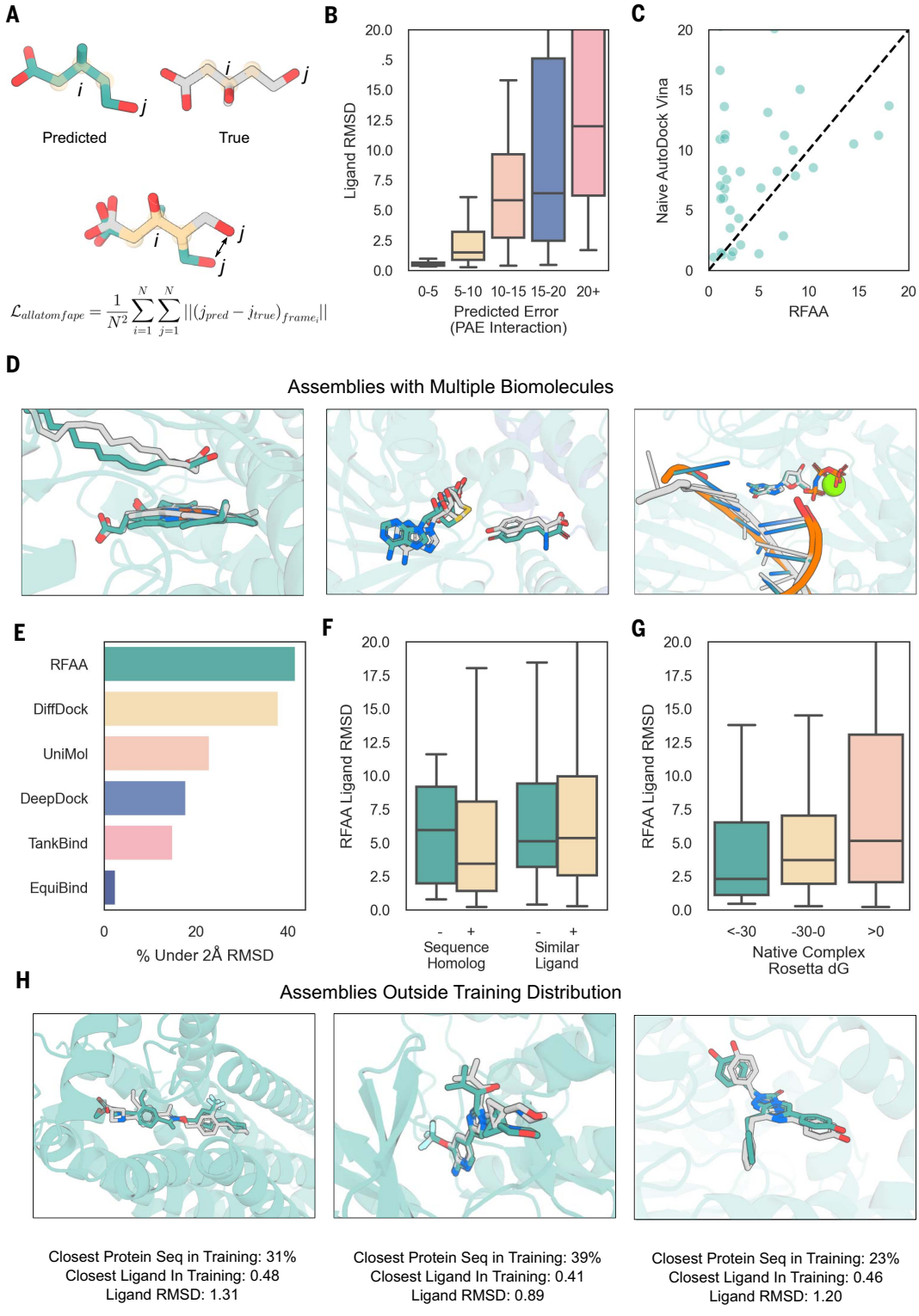


Fig. 2. RFAA can accurately predict protein–small molecule complex structures. (**A**) Every "atom" node is assigned a local coordinate frame based on the identities of its neighbors. To compute the main loss in the network, we aligned each atom's coordinate frame in the predicted and true structures and measured the error over all the other atoms. (**B**) Model accuracy correlates with error predictions. Structures and corresponding error predictions were generated for CAMEO targets (05/20/23 to 7/29/23; 261 protein–small molecule interfaces). Ligand RMSD was computed by CAMEO organizers. (**C**) RFAA outperforms AutoDock Vina on CAMEO targets (week 8/12/23 to 09/02/23; 149 protein–small molecule interfaces). Both servers had to model the protein and find pockets for all ligands present in the solved structure and the correct docks for all ligands. The ligand RMSD for both servers was computed by CAMEO organizers; the AutoDock Vina server was set up by CAMEO organizers. (**D**) Three examples of successful predictions with multiple biomolecules. Shown from left to right are fatty acid decarboxylase (PDB ID 8D8P; seq ID 34%; from CAMEO) with a heme cofactor and a lipid substrate;

a dimeric tyrosine methyltransferase (PDB ID 7UX8; seq ID 28%; CASP15 target: T1124) with an *S*-adenosyl homocysteine and tyrosine interaction; and a DNA polymerase (PDB ID 7U7W; seq ID 100%) bound to DNA, a nucleotide, and a metal ion (*31*, *66*, *67*). The following color scheme is used in all panels: Predicted protein structure (aligned to native) is indicated in transparent teal, predicted ligand conformation in teal, and native ligand conformation in gray. (**E**) Comparison to other deep learning–based docking methods. In this case, each method was applied in their respective training regime. For RFAA, this meant only having sequence and minimal atomic graph inputs, whereas for other methods, this involved providing the bound crystal structure. The ligand RMSD was computed using the PoseBusters suite, and a single example present in our training set was removed for all methods that were compared. (**F**) Comparison of RFAA predictions on recently solved PDB proteins that are new compared with the training set (homolog <1 BLAST e-value, similar ligand >0.5 Tanimoto similarity). Each set is clustered based on sequence or ligand similarity, and a random cluster representative is chosen for each. (**G**) Comparison of RFAA prediction accuracy to Rosetta ΔG energy estimates for the native complex (more than 940 cases that were successfully processed by Rosetta). RFAA makes more-accurate predictions for native complexes with low Rosetta energy. (**H**) Three examples of successful predictions with low similarity to the training set. Shown from left to right are G protein–coupled S1P receptor (PDB ID 7EW1; seq ID 31%), complex of DLK bound to an inhibitor (PDB ID 8OUS; seq ID 39%), and a *Renilla* luciferase bound to an azacoelenterazine (non-native substrate; PDB ID 7QXR; seq ID 23%) (*68*–*70*). In (B), (F), and (G), boxplots cut off at 20 Å for clarity; the center line represents the median, box limits are upper and lower quartiles, and whiskers are minimum and maximum values. The color scheme is the same as that in (D).

unit given atomic graph information. To further help the network learn about general atomic interactions, we took advantage of the commonalities between atomic interactions within proteins and many of the atomic interactions between proteins and small molecules and augmented the training data by inputting portions of proteins as atoms rather than residues (a process we term atomization). We atomized randomly selected subsets of three to five contiguous residues by deleting the sequence and template features and providing instead atom, bond, and chirality information for the atoms in those residues (an alanine would be replaced by five atom tokens, one for each heavy atom). Because the atoms are still part of the polypeptide chain, we provide the relative position of the atom tokens with respect to the other residue tokens by adding an extra bond token that corresponds to an "atom-to-residue" bond and develop a positional encoding to account for atom-residue bonds (see supplementary methods). To increase prediction accuracy on biological polymers, we trained the network on protein monomer, protein complex, and protein–nucleic acid complex examples, as previously described (*10*, *11*). All examples were cropped to have 256 tokens during the initial stages of training and 375 tokens during fine-tuning. The progress of training was monitored using independent validation sets consisting of 10% of the protein sequence clusters (see table S4).

Unlike previous protein-only deep-learning architectures (*13*–*15*), RFAA can model full biomolecular systems. In the following sections, we describe the performance of RFAA on different structure-modeling tasks. We adopted the philosophy that a single model trained on all available data over all modalities would have the greatest ability to generalize and be more accessible than a series of models specialized for specific problems.

Predicting protein–small molecule complexes

To enable blind testing of RFAA prediction performance, we enrolled an RFAA server in the blind Continuous Automated Model EvaluatiOn (CAMEO) ligand-docking evaluation, which carries out predictions on all structures submitted to the PDB each week with each enrolled server and evaluates their performance (*16*–*18*). These structures can have multiple protein chains, ligands, and metal ions (for further results on metal ions, see fig. S2). Of the CAMEO targets, 43% are predicted confidently by RFAA (PAE interaction <10), and 77% of those high-confidence structures are quite accurate, with <2-Å ligand root mean square deviation (RMSD) (Fig. 2B). One of the other servers is an implementation of the leading non–deep-learning protein–small molecule docking method AutoDock Vina, developed by the CAMEO organizers, that predicts the protein structure by homology modeling (*19*–*24*), runs AutoDock to dock the small molecules, and ranks the poses using the Vina scoring function (*9*, *19*). RFAA consistently outperformed the other servers in CAMEO on protein–small molecule modeling; for example, on cases modeled by both the RFAA and AutoDock Vina servers, RFAA modeled 32% of cases successfully (<2-Å ligand RMSD) compared with 8% for the Vina server (Fig. 2C; the Vina performance by an expert would likely be considerably improved because of the complexities of fully automatic multiple-step modeling pipelines). The most common RFAA failure mode is the placement of small molecules in the correct pockets but not in the correct orientation (fig. S3; for a further exploration of failure modes, see supplementary methods).

There were no other deep-learning docking methods (*5*, *25*–*29*) enrolled in CAMEO, but we could instead compare performance on a set of PDB structures that were solved after our training-set date cutoff (*30*) [most earlier deep-learning–based docking tools have focused on the "bound" docking problem, in which the crystal structure of the target (including side chains) are provided and hence are less well suited to CAMEO]. On this benchmark, RFAA predicts 42% of complexes successfully compared with DiffDock, which predicts 38% of complexes successfully (Fig. 2D; RFAA predicts the protein backbone and side chains in addition to the small-molecule dock, whereas DiffDock receives the crystal structure of the protein from the bound complex as input). In cases where both the bound protein structure and the pocket residues are provided, physics-based methods such as AutoDock Vina outperform RFAA (52 versus 42%), which has the much harder task of predicting both the protein backbone and side-chain details and the dock from sequence alone (fig. S4A).

To further benchmark the network, we assembled a dataset of recent PDB entries with small molecules bound that were deposited after the cutoff date for our training set and predicted full structure models for all 5421 complexes (1529 protein sequence clusters at 30% sequence identity). Although prediction performance is higher for clusters with overlap with the training set, the network also generates accurate predictions for proteins with low (BLAST e-value >1) sequence similarity to the training set (35 versus 24% success rate, respectively; Fig. 2F). We observed a similar pattern for ligand clusters (across 1310 ligand clusters); although the network makes more accurate predictions for ligands seen in training, it also can make accurate predictions on ligands that are not similar to those in training (<0.5 Tanimoto similarity; 19 versus 14% success rate) (Fig. 2F). In cases where RFAA predicts ligand placement with high confidence and RF2 has high confidence (PAE interaction <10 and pLDDT >0.8, respectively), RFAA makes higher-accuracy protein structure predictions than RF2 (fig. S3A), indicating that training with ligand context can improve overall protein-prediction accuracy. Some examples of shifts predicted by RFAA but not by RF2 include domain movements, subtle backbone movements, and flipping of side-chain rotamers to accommodate the ligand in the pocket (fig. S3, B and C).

Unlike previous methods, RFAA is able to jointly predict interactions between proteins and multiple nonprotein ligands in a single forward pass. Figure 2D shows three examples of recently solved structures with three or more components for which RFAA predictions had <2-Å ligand RMSD (when the proteins

were aligned). There are homologous complexes in the training set, so these are not de novo predictions; however, they demonstrate that RFAA can learn the multicomponent assembly prediction task. Figure 2D, right, shows a prediction for DNA polymerase (*31*) (PDB ID 7U7W) with a bound DNA, nonhydrolyzable guanine triphosphate and magnesium ion; the network received no examples of higher-order assemblies containing proteins with both small molecules and nucleic acids during training but is likely synthesizing information from multiple related binary complexes that are in the training set.

To assess whether the network can distinguish compounds known to bind from related compounds, we compared protein–small molecule complex predictions for the PoseBusters dataset for the compound known to bind and decoy molecules, including small molecules, with the highest Tanimoto similarity in the dataset. In 75.1% of cases, the PAE interaction metric of the "decoy" complex was higher (indicating lower confidence) than the native complex (fig. S7). Direct optimization on this discrimination task would likely further improve performance.

To determine the extent to which the network is reasoning over the detailed structure of protein–small molecule interactions, we investigated the correlation between prediction accuracy and the interaction energy computed by a molecular force field. We found that predictions for protein–small molecule complexes in our recent PDB set with lower computed binding energies (Rosetta ΔG) (*32*, *33*) were more accurate (Fig. 2G; 50, 25, and 22% success rates for <−30, −30 to 0, and >0 Rosetta energy units, respectively), which suggests that the network considers the detailed interactions between the protein and small molecule (although reasoning over these interactions very differently than human-designed force fields).

Predicting structures of covalent modifications to proteins

Many essential protein functions, such as receptor signaling, immune evasion, and enzyme activity, involve covalent modifications of amino acid side chains with sugars, phosphates, lipids, and other molecules (*34*–*37*). RFAA models such modifications by treating the residue and chemical moiety as atoms (with the corresponding covalent bond to the atom token in the residue) and the rest of the protein structure as residues (Fig. 3A). Unnatural amino acids can be modeled in the same way.

We benchmarked the performance of RFAA on covalent modification structure prediction on 931 recent entries in the PDB (after May 2020) and found that the network made accurate predictions (modification RMSD <2.5 Å) in 46% of cases (modification RMSD is the RMSD of the modified residue and chemical modification when the rest of the protein is aligned). As in the protein–small molecule complex case, confident predictions tend to be more accurate: 60% of structures are predicted with high confidence (PAE interaction <10), and 63% of those predictions are accurate (<2.5-Å modification RMSD) (Fig. 3B). Although the network makes slightly more-accurate predictions on cases with sequence similarity (>25% identity) to proteins in the training set, there are still many cases (27.5%) that do not have sequence overlap to the training set that are predicted with high accuracy (Fig. 3C). RFAA models interactions with covalently bound cofactors and covalently bound drugs with median RMSDs of 0.99 and 2.8 Å, respectively (Fig. 3, D and E).

Prediction of glycan structure has applications in therapeutics, vaccines, and diagnostics (*38*–*40*). RFAA can accurately model carbohydrate groups introduced by glycosylation with a median RMSD over our test set of 3.2 Å (Fig. 3D). RFAA successfully predicts glycan conformations on the *N*-acetylglucosamine-1-phosphotransferase (GNPT) gamma subunit (PDB ID 7S69) and human sperm TMEM95 ectodomain (PDB ID 7UX0), which have low sequence homology (<30%) to the RFAA training set (Fig. 3F) and have multiple monosaccharides and different branching patterns (*41*, *42*). RFAA is not simply learning how structure-building programs model glycans because the predictions match the experimental density maps (fig. S8C). The network was able to make accurate predictions of glycan interactions even when the sequences were distant from the sequences in the training set and on glycans with chains of up to seven monosaccharides (fig. S8).

It is difficult to compare RFAA with other methods because, to our knowledge, previous deep learning–based tools do not model covalent modifications to proteins. Accurate and robust modeling of covalent modifications in predicted structures should contribute to the understanding of biological function and mechanism.

De novo small-molecule binder design

Previous work on the design of small-molecule binding proteins has involved docking molecules into large sets of native or expert-curated protein scaffold structures (*43*, *44*). Diffusion-based methods can generate proteins in the context of a protein target that bind with considerable affinity and specificity (*45*) and can be trained to explicitly condition on structural features (*46*). However, present deep learning–based generative approaches do not explicitly model protein-ligand interactions, so they are not directly applicable to the small-molecular-binder design problem [in RFdiffusion, a heuristic attractive-repulsive potential encouraged the formation of pockets with shape complementarity to a target molecule, but the approach was unable to model the details of protein–small molecule interactions (*45*)]. A general method that can generate protein structures around small molecules and other nonprotein targets to maximize favorable interactions could be broadly useful.

We reasoned that RFAA could enable protein design in the context of nonprotein biomolecules after fine-tuning on structure denoising. We developed a diffusion model, RFdiffusion All-Atom (RFdiffusionAA), by training a denoising diffusion probabilistic model initialized with the RFAA structure-prediction weights to denoise corrupted protein structures conditioned on the small molecule and other biomolecular context (Fig. 4A). Input structures from the protein–small molecule dataset described above were noised through progressive addition of 3D Gaussian noise to the Cα coordinates and Brownian motion on the manifold of rotations, and the model was trained to remove this noise. In contrast to training for the unconditional generation problem and incorporating conditional information through forms of guidance (*47*, *48*), this training procedure results in an explicitly conditional model that learns the distribution of proteins conditioned on biomolecular substructure. To enable the inclusion of specific protein functional motifs when desired, we also trained the network to scaffold a variety of discontiguous protein motifs both in the presence and absence of small molecules. To generate proteins, we initialized a Gaussian distribution of residue frames with randomized rotations around a fixed small-molecule motif; at each denoising step t, we predicted the fully denoised X_0 state and then updated all residue coordinates and orientations by taking a step toward this conformation while adding noise to match the distribution for X_{t-1}. As with RFdiffusion, we investigated the use of auxiliary potentials to influence trajectories to make more contacts between small molecules and binders but found these to be unnecessary (see fig. S10C).

We evaluated RFdiffusionAA in silico by generating protein structures in the context of four diverse small molecules. Starting from random residue distributions surrounding each of the small molecules, iterative denoising yielded coherent protein backbones with pockets complementary to the small-molecule target. After sequence design using LigandMPNN (*49*, *50*), Rosetta GALigandDock (*32*) energy calculations were used to evaluate the protein–small molecule interface and AF2 predictions to evaluate the extent to which the sequence encodes the designed structure (*45*, *51*). The computed binding energies of RFdiffusionAA designs are far better ($p < 1.56 \times 10^{-12}$) than those obtained using a heuristic attractive-repulsive potential with protein-only RFdiffusion (fig. S10C). RFdiffusionAA generated backbones that could be repredicted with AF2 with backbone RMSD <2 Å for all four design cases (fig. S10C). For

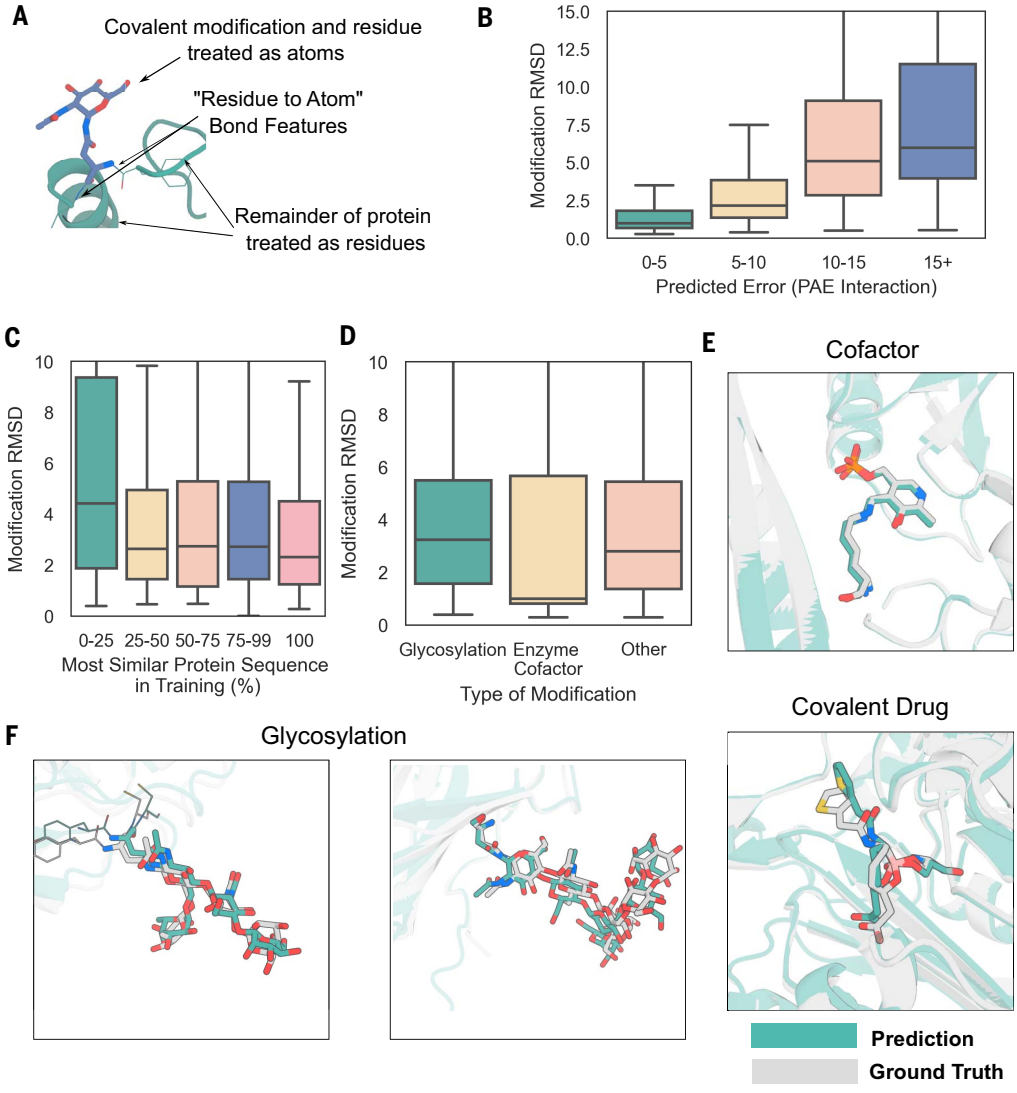


Fig. 3. Accurate prediction of protein covalent modifications. (**A**) Schematic describing how RFAA models covalent modifications to proteins. The chemical moiety that modifies the residue and the residue are modeled as atom nodes, and the rest of the protein is modeled as residues (with multiple sequence alignment and template inputs). (**B**) Model accuracy correlates with predicted error on a set of 938 recently solved structures with covalent modifications. Modification RMSD was computed by aligning the protein structure within 10 Å and computing RMSD over the modified residue and chemical modification. The boxplot is cut off at 15 Å for clarity. (**C**) Comparison of sequence identity to the training set and model accuracy. Models are generally accurate even with low sequence homology to the training set. (**D**) Comparison of model accuracy for different types of covalent modifications. (**E**) Shown at the top is an example of a successfully predicted covalently linked enzyme cofactor (PDB ID 7P3T; seq ID 28%), which is a structure of an (*R*)-selective amine transaminase. Shown at the bottom is an example of a covalently bound drug candidate (PDB ID 7TI1; seq ID 27%), which is a β-lactamase enzyme bound to cyclic boronic acid inhibitor (*71*, *72*). (**F**) Accurate predictions of glycans on the *N*-acetylglucosamine-1-phosphotransferase (GNPT) gamma subunit (PDB ID 7S69; no BLAST hits) (right) and human sperm TMEM95 ectodomain (PDB ID 7UX0; seq ID 26%) (left) (*41*, *42*). For the boxplots in (B) to (D), the center line represents the median, box limits are upper and lower quartiles, and whiskers are minimum and maximum values. In (E) and (F), predicted protein structure is indicated in transparent teal, native structure in transparent gray, predicted covalent modification in teal, and native covalent modification in gray.

each small molecule, RFdiffusionAA generates diverse protein structural solutions to the binding problem that differ from native binders to these ligands (figs. S11 and S12).

Experimental characterization of designed binders

To experimentally evaluate RFdiffusionAA across a range of design scenarios, we designed binders for three diverse small molecules: one with no protein motif included in the design parameters, one with a single-residue protein motif, and one with a four-residue protein motif (Fig. 4). We produced the proteins in *Escherichia coli* and measured ligand binding experimentally.

Digoxigenin (DIG) is the aglycone of digoxin, a small molecule used to treat heart diseases with a narrow therapeutic window (*52*), and digoxigenin-binding proteins could help reduce toxicity (*53*). Previous attempts to design digoxigenin-binding proteins relied on protein scaffolds with experimentally determined structures and prespecified binding pockets and interacting motifs (*54*). We used RFdiffusionAA to design digoxigenin-binding backbones without any prior assumption about the protein-ligand interface or backbone structure (Fig. 4A). Sequences were obtained using LigandMPNN and Rosetta FastRelax (*55*), and 4416 designs were selected based on consistency with AF2

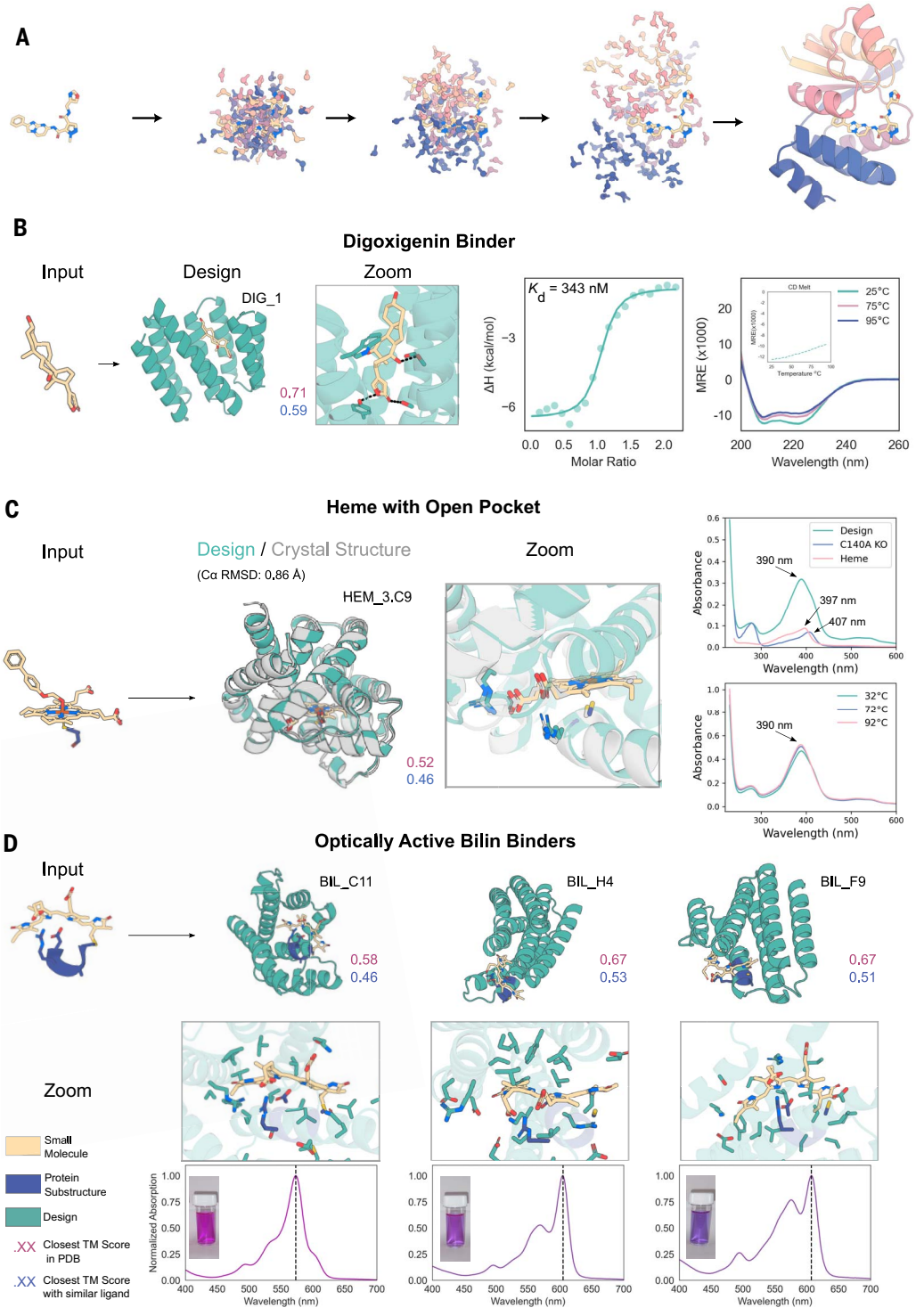


Fig. 4. Experimental characterization of binders designed with RFdiffusionAA. The following color scheme is used in all panels: The input ligand is indicated in yellow, input protein motif in blue, and diffused protein in teal; purple text indicates the closest TM score to any protein in the training set, and blue text indicates the closest TM score to any protein in the training set that has a similar ligand bound (Tanimoto similarity >0.5). (**A**) Schematic depicting the random initialization of residues surrounding a small molecule and progressive denoising by RFdiffusionAA. (**B**) Characterization of dioxigenin binder design. Shown from left to right are the input motif to RFdiffusionAA, the designed protein, a zoomed-in view of the binding-site side chains, isothermal calorimetry (ITC) measuring binding affinity (K_d = 343 nM), and a circular dichroism (CD) trace (26 μM protein concentration; the inset is a CD melt showing intensity at 220 nm across a broad range of temperatures). ΔH, enthalpy of binding; MRE, molar ellipticity. (**C**) Characterization of heme binding designs. Shown from left to right are the input motif to RFdiffusionAA, the designed protein aligned to its crystal structure (PDB ID 8VC8), a zoomed-in view of the binding site, and UV-Vis spectra showing that the designed protein matches the expected spectra for penta-coordinated heme and that mutating cysteine to alanine abolishes binding (top) and that the designed protein retains heme binding at temperatures up to 90°C (bottom). (**D**) Characterization of bilin binding designs. Row 1, left to right, shows the input motif to RFdiffusionAA and three designs with different predicted structural topologies. Row 2, left to right, shows a zoomed-in view of binding sites for each design. Row 3, left to right, shows normalized absorption spectra for the three designs. The designs have a range of maximum absorption wavelengths and hence different colors in solution (insets).

predictions and Rosetta metrics (see supplementary methods). Experimental characterization identified several DIG-binding proteins (figs. S29 and S30 and supplementary methods); the highest-affinity binder has a 343-nM dissociation constant (K_d) for free digoxigenin (as measured by isothermal titration calorimetry; Fig. 4B) and is stable at temperatures up to 95°C.

Heme is a cofactor for a wide range of oxidation reactions and oxygen transport (cytochrome P450 and hemoglobin are two notable examples), with catalytic function enabled by pentacoordinate iron binding and an open substrate pocket (*56*, *57*). Designed heme-binding proteins with these features have considerable potential as a platform for the development of new enzymes (*58*). We diffused proteins around heme with the central iron coordinated by a cysteine and a placeholder molecule just above the porphyrin ring to keep the axial heme binding site open for potential substrate molecules. Of 168 designs selected based on AF2-predicted confidence (pLDDT), backbone RMSD to design, and RMSD of the predicted cysteine rotamer to the design, 135 were well expressed in *E. coli*, and 90 had

ultraviolet-visible (UV-Vis) spectra consistent with Cys-bound heme (as judged by the Soret maximum wavelength after in vitro heme loading) (*59*). We further purified 40 of the designs and found that 33 were monomeric and retained heme binding through size exclusion chromatography. For 26 of the designs, we mutated the putative heme-coordinating cysteine residue to alanine, which led to a notable change in the Soret features in all cases (Fig. 4 and figs. S13 to S16). Twenty designs exhibit high thermostability, retaining their heme binding at temperatures above 85°C, and do not unfold at temperatures up to 95°C (Fig. 4C and figs. S13 to S16). We solved the crystal structure of heme-loaded design HEM_3.C9 to 1.8-Å resolution (PDB ID 8VC8) and found it to closely match that of the design model (0.86-Å Cα RMSD). The crystal structure verifies that heme is bound through Cys ligation in a pentacoordinate fashion with an open distal pocket (in agreement with spectroscopic data) and is further held in place with hydrogen bonds to two arginines, as designed (fig. S17).

Bilins are brilliantly colored pigments that play important roles across diverse biological kingdoms. When bilins are constrained by protein scaffolds, such as phycobiliproteins in the megadalton phycobilisome antenna complexes of cyanobacteria and some algae (*60*), their absorption features narrow, their extinction coefficients increase, and their fluorescence is markedly enhanced. We sampled diffusion trajectories conditioned on the structure of a bilin molecule attached to a four-residue peptide motif recognized by the CpcEF bilin lyase (*61*, *62*). We evaluated 94 designs with a whole-cell screen using phycoerythrobilin (PEB) as the chromophore and, on the basis of pigmentation of fluorescence, identified nine proteins dissimilar to each other and to CpcA (fig. S18A) that bind bilin (a 9.6% hit rate). We purified three designs—BIL_C11, BIL_H4, and BIL_F9—with absorption maxima at 573, 605, and 607 nm, respectively, compared with 557 nm for the CpcA-PEB [Fig. 4D and fig. S8B; the extent of red shifting correlates with computed electrostatic potential around the chromophore (fig. S19)]. Conformationally restricted bilins typically display higher fluorescence yields; absolute fluorescent yields for the BIL_C11, BIL_H4, and BIL_F9 designs are 38, 11, and 25%, respectively, based on an earlier determination of the absolute fluorescence quantum yield for CpcA-PEB of 67% (*63*) (fig. S18C). These values are much higher than those obtained previously with maquette scaffolds [fluorescence quantum yield (FΦ) values of 2 to 3%], which displayed limited bilin incorporation and less pronounced spectral enhancements (*64*). The strong coloration, absorption, and emission for these designs were absent from control *E. coli* strains that synthesize only (i) the PEB bilin and the CpcE/F lyase or (ii) PEB, CpcE/F, and maltose-binding protein (fig. S20). The 34- and 30-nm ranges in absorption and emission, respectively, covered by just one design round using a single chromophore raises the exciting prospect of tailoring the spectral profiles of designed biliproteins by manipulating the conformational flexibility of the bilin and the protein microenvironment. De novo–designed antenna complexes could harvest light over a wider range of the UV-Vis spectrum to enhance photosynthetic energy capture and conversion (*65*), and fluorescent reporter probes with tunable excitation and emission maxima would be useful biochemical tools.

The experimental validation of digoxigenin-, heme-, and bilin-binding proteins demonstrates that RFdiffusionAA can readily generate proteins with custom binding pockets for diverse small molecules. Unlike prior methods that rely on redesigning existing scaffolds, RFdiffusionAA builds proteins from scratch around the target compound, resulting in high shape complementarity in the binding pockets and reducing the need for expert knowledge. The ability of RFdiffusionAA to generalize is highlighted by the sequence and structural dissimilarity between the designs and proteins in the PDB that bind related molecules (Tanimoto similarity >0.5); the most similar protein in the PDB that binds a related molecule has a template modeling score (TM score) of 0.59 for the highest-affinity digoxigenin binder, less than 0.62 for all the characterized heme binders, and less than 0.52 for the bilin binders (fig. S21). In all cases, there is no detectable sequence similarity to any known protein.

Discussion

RFAA demonstrates that a single neural network can be trained to accurately model a wide range of general biomolecular assemblies that contain a wide diversity of nonprotein components. RFAA can make high-accuracy predictions on protein–small molecule complexes, with 32% of CAMEO targets predicted under 2-Å RMSD, and for covalent modifications to proteins, predicting 46% of recently solved covalent modifications under 2.5-Å RMSD; it can also generate accurate models for complexes of proteins with two or more nonprotein molecules (small molecules, metals, nucleic acids, etc.). Training on more-extensive datasets and/or architectural improvements will likely be necessary to generate consistently accurate predictions for protein–small molecule complexes that are on par with the accuracy of predictions that deep networks can achieve on protein systems alone. The new prediction capabilities do not come at the expense of performance on the classic protein structure prediction problem: RFAA achieves a protein-structure prediction accuracy similar to that of AF2 [median global distance test (GDT) of 85 versus 86] and a protein–nucleic acid complex accuracy similar to that of RFNA (median allatom-LDDT of 0.74 versus 0.78) (fig. S22).

Our prediction and design results suggest that RFAA has learned detailed features of protein–small molecule complexes. First, the network is able to make high-accuracy predictions for protein sequences and ligands that differ considerably from those in the training dataset (Figs. 2F and 3C), and prediction accuracy is higher for complexes with more-favorable computed interaction energies using the Rosetta physically based model (Fig. 2G). Second, our RFdiffusionAA-generated bilin, heme, and digoxigenin binders have very different structures than proteins that bind these compounds that are found in the PDB. RFAA should be immediately useful for modeling protein–small molecule complexes, in particular, multicomponent biomolecular assemblies for which there are few or no alternative methods available, and for designing small-molecule binding proteins and sensors.

Methods summary

A detailed description of dataset curation, modeling of biological inputs, data pipeline, RFAA architecture, training, in silico design methods, and experimental validation can be found in the supplementary materials.

REFERENCES AND NOTES

1. J. Jumper *et al*., Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
2. M. Baek *et al*., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi: 10.1126/science.abj8754; pmid: 34282049
3. R. A. Friesner *et al*., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004). doi: 10.1021/jm0306430; pmid: 15027865
4. M. L. Hekkelman, I. de Vries, R. P. Joosten, A. Perrakis, AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023). doi: 10.1038/s41592-022-01685-y; pmid: 36424442
5. G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, DiffDock: Diffusion steps, twists, and turns for molecular docking. arXiv:2210.01776 [q-bio.BM] (2022).
6. R. V. Honorato, J. Roel-Touris, A. M. J. J. Bonvin, MARTINI-based protein-DNA coarse-grained HADDOCKing. *Front. Mol. Biosci.* **6**, 102 (2019). doi: 10.3389/fmolb.2019.00102; pmid: 31632986
7. M. Holcomb, Y.-T. Chang, D. S. Goodsell, S. Forli, Evaluation of AlphaFold2 structures as docking targets. *Protein Sci.* **32**, e4530 (2023). doi: 10.1002/pro.4530; pmid: 36479776
8. A. M. Díaz-Rovira *et al*., Are deep learning structural models sufficiently accurate for virtual screening? Application of docking algorithms to AlphaFold2 predicted structures. *J. Chem. Inf. Model.* **63**, 1668–1674 (2023). doi: 10.1021/acs.jcim.2c01270; pmid: 36892986
9. J. Eberhardt, D. Santos-Martins, A. F. Tillack, S. Forli, AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021). doi: 10.1021/acs.jcim.1c00203; pmid: 34278794
10. M. Baek, R. McHugh, I. Anishchenko, D. Baker, F. DiMaio, Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. bioRxiv 2022.09.09.507333 [Preprint] (2022); https://doi.org/10.1101/2022.09.09.507333.
11. M. Baek *et al*., Efficient and accurate prediction of protein structure using RoseTTAFold2. bioRxiv 2023.05.24.542179 [Preprint] (2023); https://doi.org/10.1101/2023.05.24.542179.
12. C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, The Cambridge Structural Database. *Acta Crystallogr. B Struct. Sci.*

Cryst. Eng. Mater. **72**, 171–179 (2016). doi: 10.1107/S2052520616003954; pmid: 27048719
13. Z. Lin *et al*., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi: 10.1126/science.ade2574; pmid: 36927031
14. R. Wu *et al*., High-resolution de novo structure prediction from primary sequence. bioRxiv 2022.07.21.500999 [Preprint] (2022); https://doi.org/10.1101/2022.07.21.500999.
15. R. Evans *et al*., Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 [Preprint] (2022); https://doi.org/10.1101/2021.10.04.463034.
16. J. Haas *et al*., Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **86**, 387–398 (2018). doi: 10.1002/prot.25431; pmid: 29178137
17. J. Haas *et al*., The Protein Model Portal—A comprehensive resource for protein structure and model information. *Database* **2013**, bat031 (2013). doi: 10.1093/database/bat031; pmid: 23624946
18. J. Haas *et al*., Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins* **87**, 1378–1387 (2019). doi: 10.1002/prot.25815; pmid: 31571280
19. A. Waterhouse *et al*., SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018). doi: 10.1093/nar/gky427; pmid: 29788355
20. N. Guex, M. C. Peitsch, T. Schwede, Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **30**, S162–S173 (2009). doi: 10.1002/elps.200900140; pmid: 19517507
21. S. Bienert *et al*., The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017). doi: 10.1093/nar/gkw1132; pmid: 27899672
22. G. Studer *et al*., QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics* **36**, 1765–1771 (2020). doi: 10.1093/bioinformatics/btz828; pmid: 31697312
23. M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7**, 10480 (2017). doi: 10.1038/s41598-017-09654-8; pmid: 28874689
24. M. Varadi *et al*., AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022). doi: 10.1093/nar/gkab1061; pmid: 34791371
25. H. Stärk *et al*., EquiBind: Geometric deep learning for drug binding structure prediction. arXiv:2202.05146 [q-bio.BM] (2022).
26. W. Lu *et al*., TANKBind: Trigonometry-Aware Neural NetworKs for drug-protein binding structure prediction. bioRxiv 2022.06.06.495043 [Preprint] (2022); https://doi.org/10.1101/2022.06.06.495043.
27. Z. Liao, R. You, X. Huang, X. Yao, "DeepDock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information" in *Proceedings 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, I. Yoo, J. Bi, X. Hu, Eds. (IEEE, 2019), pp. 311–317. doi: 10.1109/BIBM47256.2019.8983365
28. Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, A. Anandkumar, State-specific protein-ligand complex structure prediction with a multi-scale deep generative model. arXiv:2209.15171 [q-bio.QM] (2022).
29. G. Zhou *et al*., Uni-Mol: A universal 3D molecular representation learning framework. ChemRxiv 10.26434/chemrxiv-2022-jjm0j [Preprint] (2022); https://doi.org/10.26434/chemrxiv-2022-jjm0j.
30. M. Buttenschoen, G. M. Morris, C. M. Deane, PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. arXiv:2308.05777 [q-bio.QM] (2023).
31. C. Chang, C. Lee Luo, Y. Gao, In crystallo observation of three metal ion promoted DNA polymerase misincorporation. *Nat. Commun.* **13**, 2346 (2022). doi: 10.1038/s41467-022-30005-3; pmid: 35487947
32. H. Park, G. Zhou, M. Baek, D. Baker, F. DiMaio, Force field optimization guided by small molecule crystal lattice data enables consistent sub-angstrom protein-ligand docking. *J. Chem. Theory Comput.* **17**, 2000–2010 (2021). doi: 10.1021/acs.jctc.0c01184; pmid: 33577321
33. R. F. Alford *et al*., The Rosetta All-Atom Energy Function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017). doi: 10.1021/acs.jctc.7b00125; pmid: 28430426
34. H. Bagdonas, C. A. Fogarty, E. Fadda, J. Agirre, The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **28**, 869–870 (2021). doi: 10.1038/s41594-021-00680-9; pmid: 34716446
35. S. Ramazi, J. Zahiri, Posttranslational modifications in proteins: Resources, tools and prediction methods. *Database* **2021**, baab012 (2021). doi: 10.1093/database/baab012; pmid: 33826699
36. C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, Glycosylation in health and disease. *Nat. Rev. Nephrol.* **15**, 346–366 (2019). doi: 10.1038/s41581-019-0129-4; pmid: 30858582
37. J. M. Lee, H. M. Hammarén, M. M. Savitski, S. H. Baek, Control of protein stability by post-translational modifications. *Nat. Commun.* **14**, 201 (2023). doi: 10.1038/s41467-023-35795-8; pmid: 36639369
38. R. J. Woods, Predicting the structures of glycans, glycoproteins, and their complexes. *Chem. Rev.* **118**, 8005–8024 (2018). doi: 10.1021/acs.chemrev.8b00032; pmid: 30091597
39. J. Adolf-Bryfogle *et al*., Growing glycans in Rosetta: Accurate de novo glycan modeling, density fitting, and rational sequon design. bioRxiv 2021.09.27.462000 [Preprint] (2021); https://doi.org/10.1101/2021.09.27.462000.
40. S. Jo, H. S. Lee, J. Skolnick, W. Im, Restricted N-glycan conformational space in the PDB and its implication in glycan structure modeling. *PLOS Comput. Biol.* **9**, e1002946 (2013). doi: 10.1371/journal.pcbi.1002946; pmid: 23516343
41. A. Gorelik, K. Illes, K. H. Bui, B. Nagar, Structures of the mannose-6-phosphate pathway enzyme, GlcNAc-1-phosphotransferase. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203518119 (2022). doi: 10.1073/pnas.2203518119; pmid: 35939698
42. S. Tang *et al*., Human sperm TMEM95 binds eggs and facilitates membrane fusion. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2207805119 (2022). doi: 10.1073/pnas.2207805119; pmid: 36161911
43. M. J. Bick *et al*., Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **6**, e28909 (2017). doi: 10.7554/eLife.28909; pmid: 28925919
44. N. F. Polizzi, W. F. DeGrado, A defined structural unit enables de novo design of small-molecule-binding proteins. *Science* **369**, 1227–1233 (2020). doi: 10.1126/science.abb8330; pmid: 32883865
45. J. L. Watson *et al*., De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). doi: 10.1038/s41586-023-06415-8; pmid: 37433327
46. B. Ni, D. L. Kaplan, M. J. Buehler, Generative design of de novo proteins based on secondary structure constraints using an attention-based diffusion model. *Chem* **9**, 1828–1849 (2023). doi: 10.1016/j.chempr.2023.03.020; pmid: 37614363
47. L. Wu, B. L. Trippe, C. A. Naesseth, D. M. Blei, J. P. Cunningham, Practical and asymptotically exact conditional sampling in diffusion models. arXiv:2306.17775 [stat.ML] (2023).
48. J. Ingraham *et al*., Illuminating protein space with a programmable generative model. bioRxiv 2022.12.01.518682 [Preprint] (2022); https://doi.org/10.1101/2022.12.01.518682.
49. J. Dauparas *et al*., Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022). doi: 10.1126/science.add2187; pmid: 36108050
50. J. Dauparas *et al*., Atomic context-conditioned protein sequence design using LigandMPNN. bioRxiv 2023.12.22.573103 [Preprint] (2023); https://doi.org/10.1101/2023.12.22.573103.
51. B. L. Trippe *et al*., Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv:2206.04119 [q-bio.BM] (2022).
52. Digitalis Investigation Group, The effect of digoxin on mortality and morbidity in patients with heart failure. *N. Engl. J. Med.* **336**, 525–533 (1997). doi: 10.1056/NEJM199702203360801; pmid: 9036306
53. R. J. Flanagan, A. L. Jones, Fab antibody fragments: Some applications in clinical toxicology. *Drug Saf.* **27**, 1115–1133 (2004). doi: 10.2165/00002018-200427140-00004; pmid: 15554746
54. C. E. Tinberg *et al*., Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013). doi: 10.1038/nature12443; pmid: 24005320
55. R. Das, D. Baker, Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008). doi: 10.1146/annurev.biochem.77.062906.171838; pmid: 18410248
56. T. L. Poulos, Heme enzyme structure and function. *Chem. Rev.* **114**, 3919–3962 (2014). doi: 10.1021/cr400415k; pmid: 24400737
57. X. Huang, J. T. Groves, Oxygen activation and radical transformations in heme proteins and metalloporphyrins. *Chem. Rev.* **118**, 2491–2553 (2018). doi: 10.1021/acs.chemrev.7b00373; pmid: 29286645
58. I. Kalvet *et al*., Design of heme enzymes with a tunable substrate binding pocket adjacent to an open metal coordination site. *J. Am. Chem. Soc.* **145**, 14307–14315 (2023). doi: 10.1021/jacs.3c02742; pmid: 37341421
59. M. Sono, J. H. Dawson, L. P. Hager, The generation of a hyperporphyrin spectrum upon thiol binding to ferric chloroperoxidase. Further evidence of endogenous thiolate ligation to the ferric enzyme. *J. Biol. Chem.* **259**, 13209–13216 (1984). doi: 10.1016/S0021-9258(18)90679-4; pmid: 6541651
60. N. Adir, S. Bar-Zvi, D. Harris, The amazing phycobilisome. *Biochim. Biophys. Acta Bioenerg.* **1861**, 148047 (2020). doi: 10.1016/j.bbabio.2019.07.002; pmid: 31306623
61. A. Marx, N. Adir, Allophycocyanin and phycocyanin crystal structures reveal facets of phycobilisome assembly. *Biochim. Biophys. Acta* **1827**, 311–318 (2013). doi: 10.1016/j.bbabio.2012.11.006; pmid: 23201474
62. C. Zhao *et al*., Structures and enzymatic mechanisms of phycobiliprotein lyases CpcE/F and PecE/F. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13170–13175 (2017). doi: 10.1073/pnas.1715495114; pmid: 29180420
63. S. F. H. Barnett *et al*., Repurposing a photosynthetic antenna protein as a super-resolution microscopy label. *Sci. Rep.* **7**, 16807 (2017). doi: 10.1038/s41598-017-16834-z; pmid: 29196704
64. J. A. Mancini *et al*., De novo synthetic biliprotein design, assembly and excitation energy transfer. *J. R. Soc. Interface* **15**, 20180021 (2018). doi: 10.1098/rsif.2018.0021; pmid: 29618529
65. A. Hitchcock *et al*., Redesigning the photosynthetic light reactions to enhance photosynthesis - the PhotoRedesign consortium. *Plant J.* **109**, 23–34 (2022). doi: 10.1111/tpj.15552; pmid: 34709696
66. K.-L. Wu *et al*., Expanding the eukaryotic genetic code with a biosynthesized 21st amino acid. *Protein Sci.* **31**, e4443 (2022). doi: 10.1002/pro.4443; pmid: 36173166
67. L. L. Rade *et al*., Dimer-assisted mechanism of (un) saturated fatty acid decarboxylation for alkene production. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2221483120 (2023). doi: 10.1073/pnas.2221483120; pmid: 37216508
68. Y. Yuan *et al*., Structures of signaling complexes of lipid receptors S1PR1 and S1PR5 reveal mechanisms of activation and drug recognition. *Cell Res.* **31**, 1263–1274 (2021). doi: 10.1038/s41422-021-00566-x; pmid: 34526663
69. K. Le *et al*., Discovery of IACS-52825, a potent and selective DLK inhibitor for treatment of chemotherapy-induced peripheral neuropathy. *J. Med. Chem.* **66**, 9954–9971 (2023). doi: 10.1021/acs.jmedchem.3c00788; pmid: 37436942
70. A. Schenkmayerova *et al*., Catalytic mechanism for Renilla-type luciferases. *Nat. Catal.* **6**, 23–38 (2023). doi: 10.1038/s41929-022-00895-z
71. E. Konia *et al*., Rational engineering of *Luminiphilus syltensis* (*R*)-selective amine transaminase for the acceptance of bulky substrates. *Chem. Commun.* **57**, 12948–12951 (2021). doi: 10.1039/D1CC04664K; pmid: 34806715
72. K. Raja Reddy *et al*., Broad-spectrum cyclic boronate β-lactamase inhibitors featuring an intramolecular prodrug for oral bioavailability. *Bioorg. Med. Chem.* **62**, 116722 (2022). doi: 10.1016/j.bmc.2022.116722; pmid: 35358864
73. R. Krishna *et al*., Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Dryad (2024); https://doi.org/10.5061/dryad.mcvdnck6v
74. R. Krishna, Generalized biomolecular modeling with RoseTTAFold All-Atom. Zenodo (2024); https://doi.org/10.5281/zenodo.10699231

ACKNOWLEDGMENTS

We thank L. Goldschmidt and K. VanWormer for maintaining the computational and wet-lab resources at the Institute for Protein Design. We thank J. Watson and D. Juergens for helpful conversations during the development of the method and for providing the figure color scheme used in this work, J. Dauparas for helpful conversations about model improvements and developing LigandMPNN, G. Zhou for help with GALigandDock, and M. Kennedy for reading earlier drafts of the manuscript. We

thank L. Milles for providing custom plasmids. We thank T. Nguyen for assistance with the size exclusion chromatography with multi-angle static light scattering experiment. Crystallographic data were collected on AMX of the National Synchrotron Light Source II, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under contract no. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS), through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). We also thank X. Robin for helping set up our CAMEO server and providing the Vina and AD4 servers for us to benchmark our results and M. Buttenschoen for providing metadata from their PoseBusters benchmark results. **Funding:** We thank Microsoft for their generous donation of Azure Compute Credits, and Perlmutter grant NERSC award BER-ERCAP0022018 for access to the Perlmutter high-performance computing resources. This work was supported by gifts from Microsoft (R.K., P.S., D.B.); the Howard Hughes Medical Institute (D.B., G.R.L.); the New Faculty Startup Fund from Seoul National University (M.B.); the Schmidt Futures program (J.W., R.M., F.D.); the Open Philanthropy Project Improving Protein Design Fund (J.W., I.K., G.R.L.); grant no. INV-010680 from the Bill and Melinda Gates Foundation (W.A.); the Audacious Project (P.V.); the Washington State General Operating Fund supporting the Institute for Protein Design (P.V.); the Defense Threat Reduction Agency (DTRA) (G.R.L.); the Washington Research Foundation's Innovation Fellows Program (G.R.L.); a Faculty of Science PhD studentship from the University of Sheffield (F.S.M.-B.); federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. 75N93022C00036 (I.A.); Amgen (I.R.H.); Juvenile Diabetes Research Foundation International (JDRF) grant no. 2-SRA-2018-605-Q-R (X.L.); Helmsley Charitable Trust Type 1 Diabetes (T1D) Program grant no. 2019PG-T1D026 (X.L.); Defense Threat Reduction Agency grant HDTRA1-19-1-0003 (X.L.); Bill and Melinda Gates Foundation no. OPP1156262 (X.L.); Synergy award 854126 from the European Research Council (G.A.S., C.N.H.); a Royal Society University Research Fellowship (award no. URF\R1\191548 to A.H.); and Human Frontiers Science Program grant RGP0061/2019 (F.D.).

Author contributions: Research design: R.K., I.A., M.B., F.D., D.B.; Development of the RFAA architecture and training regimen: R.K., J.W., F.D.; Evaluation of RFAA on different structure prediction tasks: P.S., R.K., I.R.H., R.M.; Development RFdiffusionAA: W.A.; Generation of designs for digoxigenin binders: P.V., G.R.L.; Execution of experiments for digoxigenin binders: P.V., G.R.L., D.V., X.L.; Generation of designs and execution of experiments for heme binders: I.K.; Generation of designs for bilin binders: W.A.; Execution of experiments for bilin binders: F.S.M.-B.; Contribution of code and ideas: I.A., G.A.S., .B., F.D.; Execution of crystallography experiments: A.K., E.B., A.K.B.; Supervision throughout the project: D.B., A.H., C.N.H.; Writing – original draft: R.K., J.W., W.A., D.B. Writing – review and editing: All authors. **Competing interests:** R.K., J.W., W.A., A.I., F.D., and D.B. have filed for a provisional patent covering the work presented. The other authors declare no competing interests. **Data and materials availability:** Additional data files that contain successful design models and possible training-set sequences can be found at Dryad (*73*). Code and neural network weights are available at https://github.com/baker-laboratory/RoseTTAFold-All-Atom and https://github.com/baker-laboratory/rf_diffusion_all_atom for RFAA and RFdiffusionAA, respectively. These repositories are archived at Zenodo (*74*). All other data are available in the main text or supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adl2528
Materials and Methods
Figs. S1 to S34
Tables S1 to S16
References (*75*–*120*)
MDAR Reproducibility Checklist

Submitted 9 October 2023; accepted 27 February 2024
Published online 7 March 2024
10.1126/science.adl2528